Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 2438–2447

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Real-Time Deep Learning Approach for Pedestrian Detection and Suspicious Activity Recognition

Ujwalla Gawande[a], Kamal Hajari[b], Yogesh Golhar[c]

[a] *IT Department, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India-441110*
[c]*CSE Department, St. Vincent Palloti College of Engineering, Nagpur, Maharashtra, India-441108*

Abstract

Pedestrian detection, tracking, and suspicious activity recognition have grown increasingly significant in computer vision applications in recent years as security threats have increased. Continuous monitoring of private and public areas in high-density areas is very difficult, so active video surveillance that can track pedestrian behavior in real time is required. This paper presents an innovative and robust deep learning system as well as a unique pedestrian data set that includes student behavior like as test cheating, laboratory equipment theft, student disputes, and danger situations in institutions. It is the first of its kind to provide pedestrians with a unified and stable ID annotation. Again, presented a comparative analysis of results achieved by the recent deep learning approach to pedestrian detection, tracking, and suspicious activity recognition methods on a recent benchmark dataset. Finally, paper concluded with investigation new research directions in vision-based surveillance for practitioners and research scholars.

Keywords: Pedestrian detection; Video Surveillance; Tracking; Suspicious activity.

1. Introduction

Video surveillance is now installed everywhere to track and monitor pedestrians or criminals in streets, airports, banks, prisms, laboratories, shopping centers, etc. [1]. The surveillance system is based on a closed-circuit television (CCTV) system. Recently, Pan-Tilt-Zoom (PTZ) cameras have many advantages over traditional CCTV cameras. The main advantage of a PTZ camera is that it allows users to view more content than a fixed camera. The featuresof the PTZ camera include: 1) The user can pan left and right and tilt up and down to obtain a complete 180° view,whether it is left or right or up and down. If installed and positioned correctly, advanced PTZ cameras can provide a complete 360° field of view. Therefore, a single pan/tilt camera can replace two or even three fixed-view cameras,which is very suitable and can almost eliminate most of the blind spots on cameras with deviated fixed-view angles.A PTZ camera is programmed to rotate automatically in multiple directions at a different view of an area.Researchers main focus is

10.1016/j.procs.2023.01.219

to develop a video surveillance system that can assess pedestrian methods in real time [2]. The challenge of identifying pedestrians in crowded environments becomes extremely challenging in real-time whenlow-resolution images, motion blur, contrast illumination, scale or size of pedestrian changes, and entirely orpartially obscured outlines are present. Fig.1 describes the proposed approach motivation. Pedestrian publicuniversity dataset such as INRIA [2], Caltech [1], MS COCO [3], KITTI [5], and ETH [4] datasets, pedestrian casesare typically modest. Due to restrictions such as 1) cloudy presentation, 2) confused and imprecise boundary, 3) duplicated pedestrian occurrences, 4) tiny and big dimension occurrences with distinctive properties, etc., localizing these small instances in the presence of illumination change and occlusion is a vital operation. The advanced research onpedestrians" analysis conducted on publicly available benchmark datasets.

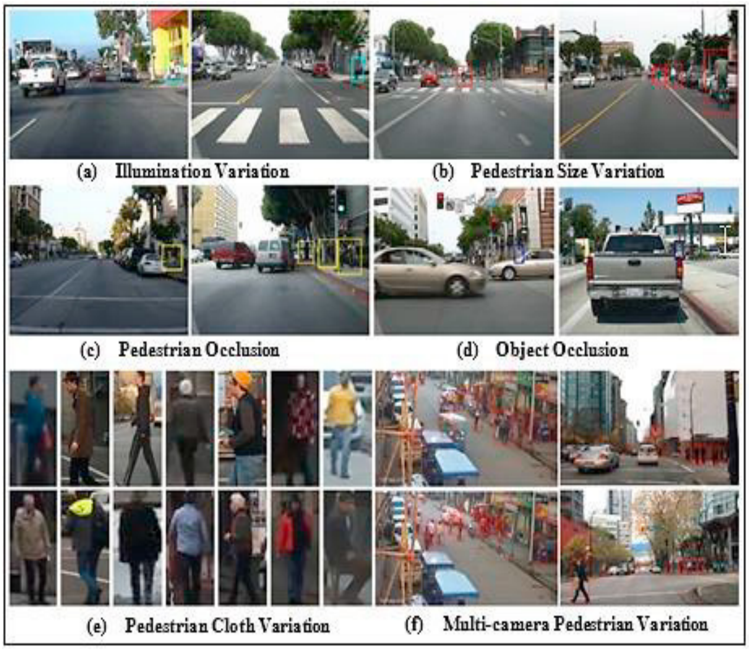
(a) Illumination Variation
(b) Pedestrian Size Variation
(c) Pedestrian Occlusion
(d) Object Occlusion
(e) Pedestrian Cloth Variation
(f) Multi-camera Pedestrian Variation

Fig.1. (a) Issues and challenges of ETH [2] and Caltech [3] datasets. (a) pedestrian significant change in the visual as the illumination changes. (b) pedestrian scale or size changes in images changed significantly. (c) pedestrian occlusion affects the detection and tracking results. (d) pedestrian occlusion with other road objects effects detection accuracy. (e) pedestrian cloth variation affects the detection algorithm accuracy. (f) Multi-camera captured direction represents a different visual appearance.

These datasets have several limitations, including: 1) a limited variety of pedestrian instances captured in a supervised pattern. 2) The size of the dataset is small and contains the least scenarios. 3) limited environment such as urban road and city only. No student behavior dataset is available. A robust and novel deep learning model and student academic environment dataset. During each sequence of frames in the video, human experts annotated the behavior of student pedestrians. It provides data in three categories. 1) Using bounding boxes to locate pedestrians. 2) fully labelled; 3) unique IDs are used as a class category for annotated pedestrians. The proposed contributions are as follows:

1. To solve existing state-of-the-art database concerns such as size and illumination variance in pedestrian images, It presents the unique enhanced Mask R-CNN deep learning architecture.
2. Student's normal and suspicious activities are recorded in the proposed dataset.
3. Within the framework of the proposed pedestrian dataset for academic settings along with comprehensive review of previous work and compare existing techniques.

The remaining research paper is categorized as follows: the most significant contribution of the new pedestrian dataset, as well as concerns and challenges in the academic context discuss in section 2. A deep learning architecture is described in section 3. The outcomes of the empirical examination are discussed in section 4. Finally, in the Section 5 the research paper ended with conclusion and research direction.

2. Related Work

This section describes the most relevant and recent pedestrian datasets. In addition, it discusses the advanced deep learning approaches to pedestrian detection, tracking, and suspicious activity recognition, along with its limitation.

2.1. State-of-the-art Pedestrian Dataset

Currently pedestrian datasets used by researchers for pedestrian detection, tracking, and suspicious activity recognition. First, the Caltech dataset contains 2,300 unique pedestrians and 350,000 annotated bounding boxes to represent these pedestrians. This dataset was created on the city road and using the camera mounted on the vehicle [3]. Second, the MIT dataset is the well-known pedestrian dataset, which consists of high-quality pedestrian sample images. It contains 709 unique pedestrians. Whether in front view or back view, the range of pose images taken on city streets [4] is relatively limited. Third, Daimler, this dataset captures people walking on the street through cameras installed on vehicles in an urban environment during the day. The data set includes pedestrian tracking attributes, annotated labelled bounding boxes, ground truth images, and floating disparity map files.

The training set contains 15560 pedestrian images and 6744 annotated pedestrian images. The test set contains 21,790 pedestrian images and 56,492 annotated images [5]. The ATCI dataset is a pedestrian database acquired by a normal car"s rear-view camera, and it"s used to test pedestrian recognition in parking lots, urban environments city streets, and private as well as public lanes. The data set contains 250 video clips, each of 76 minutes, and 200,000 marked pedestrian bounding boxes, captured in day-light scenes, with contrasting weather scenarios [6]. The ETH dataset is used to observe the traffic scene from the inside of the vehicle. The behavior of pedestrians is recorded and placed over vehicles. In an urban setting, the dataset can be used for pedestrian recognition and tracking via mobile platforms. Road cars and pedestrians, are included in the dataset [7]. The TUD-Brussels dataset was created usinga mobile platform in an urban environment.

Crowded urban street behavior was recorded vehicle embedded cameras. It can be used in car safety scenarios in urban environments [8]. One of the most abstract pedestrian detection data sets is the INRIA dataset. It incorporates human behavior, as well as a mobile camera and complex background scenes, with various variations in posture, appearance, dress, background, lighting, contrast, etc. [9]. The PASCAL Visual Object Classes (VOC) 2017 and 2007 collection contains static objects in an urban setting with various viewpoints and positions. This dataset was created with the goal of recognizing visual object classes in real-world scenarios. Animals, trees, road signs, vehicles, and people are among the 20 different categories in this collection [10]. The Common Object in Context was constructed using the MS COCO 2018 dataset [11]. (COCO). The 2018 dataset wasrecently utilized to recognize distinct things in the context while focusing on stimulus object detection.

The annotations include different examples of things connected to 80 different object categories and 91 different human segmentation categories. For pedestrian instances, there are key point annotations and five picture labels per sample image. (1) real-scene object detection with segmentation mask, (2) panoptic semantic segmentation, (3) pedestrian keypoint detection and evaluation, and (4) Dense Pose estimation in a congested scene is among the COCO 2018 dataset challenges [12].

For street picture segmentation, the Mapillary Vistas Research dataset is employed [13]. Pedestrians and other non-living categories are solved using panoramic segmentation, which successfully merges the concepts of semantic and instance segmentation. A comparison of pedestrian databases and their video surveillance purposes is shown inTable 1. In addition, we've included proposed dataset, which will be introduced in the next section. The connection is made based on the dataset's use, size, environment, label, and annotation.

2.1. Proposed Deep learning architecture and Academic Environment Pedestrian Dataset

In this section, the proposed framework from a different perspective, as captured by a high-qualityDSLR camera. The proposed video acquisition framework records the video at 30f/s along with 384x2160 resolution. The size of the dataset is 100GB. The student behavior frames shown in Fig.2. The orientation of the camera is in the range of 45^o to 90^o. Yeshwantrao Chavan College of Engineering (YCCE), Nagpur student academic activity behavior recorded in the proposed dataset. The student age is between 22-27, including both male and female. Out of which, 65% are male and 35% are female. The academic environment dataset consists of different behaviors such as lab student activities, exam hall, classroom, student cheating behavior, dispute, and stealing a mobile phone and lab electronic devices [34].

Table 1. Comparison of benchmark pedestrian dataset the pedestrian detection, tracking.

Dataset	Dataset size	Annotation	Environment	Year	Ref.	Issues and Challenges
Caltech	250000 frames	2300 unique pedestrian	City street	2012	[3]	Only urban roads are captured.
MIT	709 unique pedestrians	No annotated pedestrian	Day light scenario	2000, 2005	[4]	Missing annotation not allow userto verify different techniques.
Daimler	15,560 unique pedestrians	Ground truth withbounding boxes.	City street	2016	[5]	Only urban roads are captured.
GM-ATCI	Video clips:250	Annotated pedestrian bounding boxes: 200K	Day and complex weather and lighting	2015	[6]	Only urban roads are captured. Side view of road not captured.
ETH	Videos	Annotated cars and pedestrians	City street	2010	[7]	Small size dataset. Limited scenarios cover.
TUD Brussels	1092 frames	Pedestrian Annotation	City street	2009	[8]	Only urban roads are captured.
INRIA	498 images	Manual Annotations	City street	2005	[9]	Only urban roads are captured.
PASCAL VOC 2012	11,530 images, 20 objects classes	ROI Annotated 27,450	City street	2012	[10]	Only urban roads are captured.
MS COCO 2017	328,124 images	Segmented people object	City street	2017	[11]	Only urban roads are captured.
MS COCO 2015	328,124 images	Segmented people object	City street	2015	[12]	Only urban roads are captured.
Mapillary Vistas dataset 2017	152 obj. , 25300 img.	Instance segmentation	City street	2017	[13]	Only urban roads are captured. Side view of road not captured.

At the frame level, domain experts annotate the pedestrian video sequence. The labelling stage contains three phases: 1) human identification. 2) tracking, and 3) detection of suspicious activities. First, Mask R-CNN [12] method was used to determine the location of the pedestrian in the frame, followed by manual validation and correction of the data. Next, a deep sort [14] model was used for extracting tracking information. At last, with these two basic operations, get a rectangle bounding box around pedestrians that defines the ROI foreach human. The last stage of the updating process is performed manually, with human expert knowledge in the academic environment. Height, age, enclosing box, unique Id, feet, frame, body size, hairstyle, hair color, head attachments, clothing, mustache stubble activities, and accessories are all given for each human instance in the frame mostly on the label.

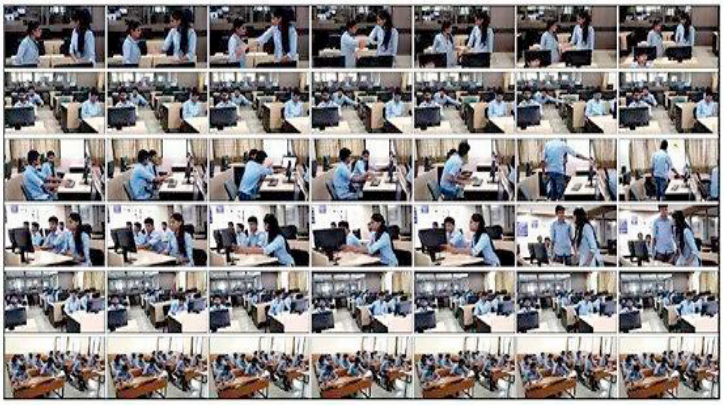

Fig.2: An example of the designed database. Lab fight between two girls – 1st Row. The scenario of snatching the phone is depicted in the 2nd row. A scenario of a student threatening is depicted in the 3rd row. The 4th row describes thesame critical situation. The 5th row depicts a situation in which students steal lab material. The sixth row depicts exam cheating scenario in examination hall.

3. Recent Deep Learning Architecture

The current deep learning-based pedestrian detection, tracking, and suspicious activity recognition systems are not as accurate and fast as human vision [2]. Pedestrian detection, tracking, and activity recognition are now separated into two categories: CNN and deep learning. V. Jones [3] approach used in pedestrian detection for face recognition. Again, HOG [5] and DPM [4] conventional approaches are used for pedestrian detection. These procedures are computationally intensive and time-consuming, and they necessitate the participation of humans. CNN-based deep learning techniques have grown in prominence as a result of their accuracy in pedestrian identification [7,8]. R-CNN [9] is the first deep learning model for object detection. Multiple stage convolutional network such Mask R-CNN [9], R-CNN family other variant as (Fast and Faster R-CNN model) [9][10][11][12]. Other, CNN models having single stages such as You Only Look Once (YOLO) [14] and SSD [15] are examples of deep learning approaches.As a result, real-time pedestrian detection is now unsuitable. As a result, Redmon et al. [15] introduced the YOLO net, that is an object regression architecture, to increase detection speed and accuracy. The proposed improved YOLOv5 method effectively detects small and constant pedestrians.

2.1. YOLOv5 Deep Learning Architecture

The YOLOv5 detector has only one stage. The YOLOv5 architecture contains three sections. 1) A solid foundation. 2) The output, and 3) the neck. The input picture features are extracted first by the backbone portion. For scale invariant feature extraction, CNN and max-pooling backbone networks are used [29]. The feature map development process is divided into four tiers in the backbone network. Each layer generates a feature map with the following dimensions: 152x152 pixels, 76x76 pixel resolution, 38x38 pixel resolution, and 19x19 pixel resolution. The neck network integrates feature maps of several levels to capture additional contextual information and prevent information loss. For multi-scale features, a recursive neural network is created, and pixel grouping backbone networks are employed for feature engineering. In a top-down method, semantic features are provided via a feature pyramid network. The bottom-up approach uses a pixel aggregation network for object localization. In the neck network, it can see three feature fusion components of different scales with sizes of 76x76x255, 38x38x255, and 19x19x255, where 255 is the network's image intensity range. The CSP network aims to improve inference speed. In the neck, the CSP network replaces the leftover units with CBL modules. The SPP module combines the benefits of the largest pooling with the flexibility of varying kernel sizes. The feature map in the input is mostly compressed. It compresses the extracted feature, resulting in a considerable reduction in feature extraction time. Again, it compresses features and removes the most important ones. Following that, it went through the intricacies of the upgraded YOLOv5 architecture.

2.1. Improved Mask R-CNN Deep Learning Architecture

To detect pedestrians on several scales, by leveraging scale-independent convolutional feature construction, the suggested Improved Mask R-CNN addresses the existing approach difficulties. Fig.3 depicts the basic concept of scale-independent feature map generation. A unique Improved R-CNN framework based on the Faster R-CNN pipeline [12] has been proposed as a result of the aforesaid proposal. The proposed Improved R-CNN is a unified architecture that combines a scale-independent feature map with a two-stage backbone network.

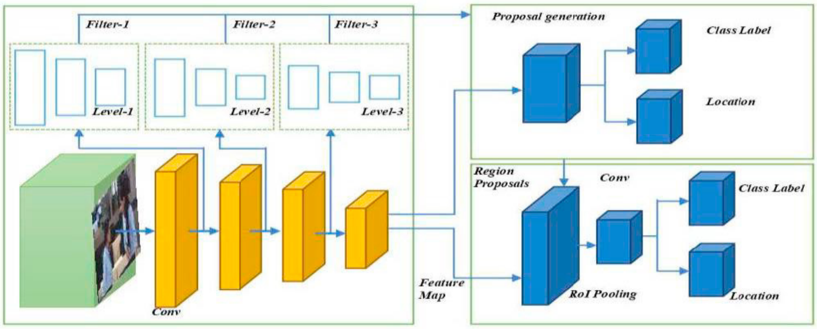

Fig.3. Depicts the basic concept of scale independent feature map generation.

The Improved R-CNN, as illustrated in Fig.4, takes input image and runs it through the common convolutional layers to obtain its whole local features. The scale-independent feature map and multiple backbone structure, that is useful for the present input about certain scales, may always help to increase the decisive outcomes. As a result, improved R-CNN can outperform traditional R-CNN detection across a large number of input scales. Improved R-CNN is also particularly efficient and effective of training and testing time because it combines convolutional characteristics for the input image. The traditional Mask R-CNN uses a binary convolution mask for every candidate region.

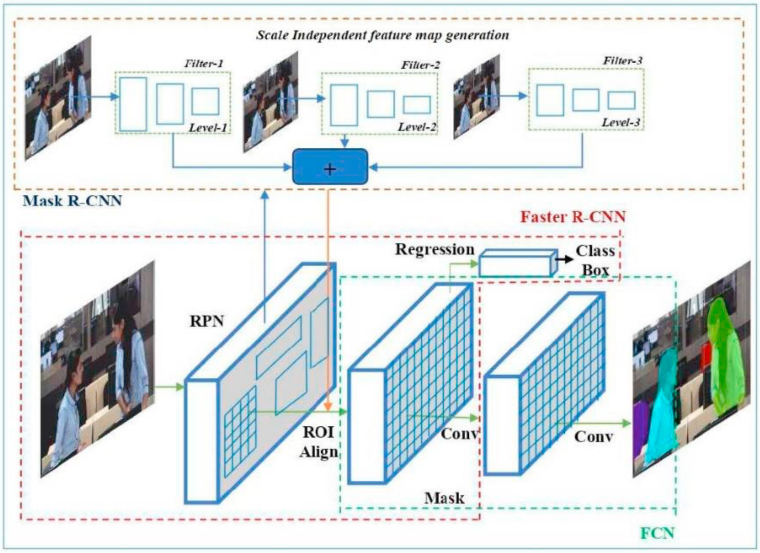

Fig.4. Improved Mask R-CNN Architecture.

A normalized framework is used to train the Scale invariant modified Mask R-CNN on the Microsoft COCO pedestrian dataset. Its testing section consisted of two primary steps: proposal region construction and pedestrian classification. The RoIs generated by the proposed region generation process might not even contain the requested object. The Region of Interest are categorized as an object or the background in the classification step. The traditional R-CNN network is very expensive for the original image to the network, notwithstanding its remarkable outcomes in general of object detection accuracy. The RoIs must be generated using the proposal region generation technique, which takes time. Since the exponential function for the items is unavailable, the tiniest objects are not categorized efficiently. Mask R-CNNs in a practical utility for pedestrian monitoring are limited due to these two flaws. In the traditional Mask R-CNN, the human examples, which differ in scale, are not recognized adequately. As a result, this problem must be addressed by creating scale-independent extracted features for the various scales of human instances. To construct the probability value, the multi resolution images were mixed with various sizes of anchor box. This score is then combined with the feature map created previously in the process. Because of the scale-independent feature map, the region proposal accuracy improves. The steps in creating a scale-independent map for object detection discuss in brief as follows:

Algorithm: Features extraction algorithm for generation Scale-independent Object detection

```
Input: MultiScaleFeature, scale invariant mask, MultiScaleImage, multi-scale image,
Output: scaleFeatureMatrix, the scale-independent matrix
Read all the image at multiple scale.
for $scalefeatureMatrixf_i \leftarrow 1$ to $K$ do
    Multi-scale image convolution process
    for $MultiScalFj \leftarrow 1$ to $T$ do
        Multi-scaled invariant mask generated and convolved
        CScoreConv (MultiScaleImage, MultiScaleFeature)
        if CScoreConv = null then
            exit out of iteration
        else
            $CuScore \leftarrow CuScore + CuScore$
        end
    end
    $scaleFeatureMatrix \leftarrow CuScore$
end
```

To identify the varied scale pedestrian efficiently, scale-independent local features are combined with other convolution layers. Following that, the RoI align procedure is used to align each region's suggestions. Other Mask R-CNN processes are used in subsequent processing. The Mask R-CNN architecture for pedestrian detection has been discussed. Finally, the bounding box and segmented mask are used to represent all of the discovered and finally, results are represented using class annotation for each object in the scene.

4. Result and Discussion

In this section, the results of three tasks performed using methods regarded to be cutting-edge technology as pedestrian detection, tracking, and suspicious activity recognition. It also presents the results acquired using such strategies in the academic environment database, it also presents baseline findings usingthe same technique in a well-known dataset. Pedestrian Detection is the first step. Both the R-FCN [15] and RetinaNet [16], Mask R-CNN [17] deep learning frameworks excelled in the PASCAL VOC [17] problems,particularly in pedestrian identification.

4.1 Performance evaluation on the state-of-the-art Pedestrian Detection

The R-FCN [15] and RetinaNet [16], Mask R-CNN [17] computational intelligence system provides the benchmark performance for pedestrian detection given that both performed very well in the PASCAL VOC [17] challenges, particularly in the pedestrian detection issues. It compared the predicted dataset performance of the two methods with the results seen in the PASCAL VOC 2007 and 2012 datasets. On top of the ResNet, RetinaNet leverages the Feature Pyramid Network (FPN) as its support system. Variations in position are encoded using a particular convolutional layer by R-FCN [15]. Instead of a completely linked layer, it makes use of ROI max pooling. Again, try Mask R-CNN on the suggested dataset. The dataset is divided into three categories: training (60%) and testing (20%) for real-time queries. The experimental findings for the proposed and current methodologies are presented in Table 2. AP_{IoU}=0:5

represents the Average Precision (AP) at the Intersection of Union (IoU) values of the common evaluation measure with the value set to 0.5.

Table 2. Comparative analysis of proposed method and existing method on Proposed dataset And PASCAL VOC 2007

Methodology	**Backbone Structure**	**PASCAL VOC Dataset**	**Proposed Dataset**
R-FCN CNN Net [15]	ResNet-101 Network	84.43 ± 1.85	59.29 ± 1.31
RetinaNet CNN Net [16]	ResNet-50 Network	86.44 ± 1.03	63.10 ± 1.64
Proposed Mask R-CNN	ResNet-101 Network	87.41 ± 1.02	65.10 ± 1.44

4.2 Performance evaluation on the state-of-the-art human Tracking in Surveillance System

The Tracktor CV [2] and V-IOU [18] techniques give the state-of-the-art, for two reasons: 1) best performer in the MOT challenge; and 2) open source pre-builded framework. The two phases of the TracktorCV approach are as follows: 1) a regression component that uses the output of the detector stage to modify the bounding box's current location; and 2) a detecting component that keeps the set of frames for the subsequent frames. For both methodologies, it is seen that there is a positive association between failures that are connected to crowds and two worrying instances: 1) scenarios where trajectories intersect individuals at every second due to dense pedestrian congestion; and 2) when crucial deformations of the person silhouettes occur. The proposed database comprises of more intricate pictures with dense backdrops, including several situations with organization. It provides an overview of the findings in Table 3.

Table 3. Performance comparison of the two cutting-edge tracking methods using the suggested dataset and MOT datasets.

Methodology	Backbone	MOTA Measure	MOTP Measure	F1-Measure
TracktorCv [2]	MOT-17	65.20 ± 9.60	62.30 ± 11.00	89.60 ± 2.80
	Proposed dataset	56.00 ± 3.70	55.90 ± 2.60	87.40 ± 2.00
V-IOU [18]	MOT-17	52.50 ± 8.80	57.50 ± 9.50	86.50 ± 1.90
	Proposed dataset	47.90 ± 5.10	51.10 ± 5.80	83.30 ± 8.40
Proposed Mask R-CNN	MOT-17	67.50 ± 9.80	59.50 ± 10.50	89.50 ± 2.87
	Proposed dataset	70.90 ± 6.12	57.10 ± 15.80	83.10 ± 18.39

4.3 Performance evaluation on the state-of-the-art Pedestrian Suspicious Activity Recognition

In surveillance videos, abnormal activities are detected by varying object behaviors in scenes with varying appearances, scales, lighting conditions, and occluded trajectories. In a crowd area, detecting individual pedestrians is an important process. The aforementioned techniques are not suitable in these circumstances. The use of motion has been the focus of other recent investigation groups [20][21]. The extraction of local spatiotemporal cuboids from optical flow or gradient patterns has also been attempted [23]. Challa S.K. et al. [24] presented a multi-branch CNN-BiLSTM model for human activity recognition using wearable sensor data. An activity recognition algorithm based on CNN filters is used in this approach. Next, Jain R. et al. [25] proposed a deep ensemble learning approach for lower extremity activities recognition using wearable sensors. Semwal V. B. and colleagues [26] proposed an improved method for the selection of features for the recognition of human walking activities using bio-geography optimization. An ensemble learning approach was used by Semwal et al. [27] in order to create an optimized hybrid deep learning model for recognizing human walking activities. Other handcrafted features were considered by another author. An invariant gait recognition-based person identification method was presented by Semwal et al. [28]. According to Bijalwan et al. [29], multi-sensor based biomechanical gait analysis can be carried out using wearable sensors combined with vision. Semwal et al. [30] presented a pattern identification of different human joints for different human walking styles using an inertial measurement unit (IMU) sensor. Dua N. et al. [31] proposed a multi-input CNN-GRU based human activity recognition using wearable sensors. Bijalwan V. et al. [32] proposed a heterogeneous Computing Model for Post-injury Walking Pattern restoration and Postural Stability Re-habilitation Exercise Recognition". Again, although the above methods have proved their effectiveness in experiments, most of them only cover the detection of abnormal activities in local or global areas. As shown in Table 4, joint deliberation of motion flows pattern, varying size of objects, and interactions between adjacent objects can be used to represent pedestrian activities in a high-density scene and enhance the performance of unusual activity detection.

Table 4. Comparative analysis of frame level Suspicious activity recognition using Equal Error Rate (EER).

Methodology/Technique	Pedestrian 1 (%)	Pedestrian 2 (%)	Average (%)
Social Force Map [22]	36.5%	35.0%	35.7%
MDT-spatial [23]	32%	38%	34%
Multibranch CNN-BiLSTM model [24]	44%	26%	32%
Lower extremity activities recognition [25]	45%	27%	35%
Optimized feature selection [26]	46%	29%	38%
Hybrid deep learning model [27]	48%	25%	34%
Pose Invariant Gait [28]	45.2%	29%	33%
Multi-sensor Gait [29]	45%	31.2%	39%
Human joints using IMU [30]	41%	27%	35%
Multi-input CNN-GRU [31]	38%	37%	37%
Heterogeneous Computing Model [32]	47%	31%	40%

5. Conclusion

In this paper, the academic environment database is proposed, which comprises video sequences of pedestrians in indoor academic environments that are annotated at the frame level. The pedestrian database contains the behavior of students in the institution. This is the first of its kind dataset that provides a unified and stable pedestrian ID annotation, making it suitable for pedestrian detection, tracking, and behaviour detection. It also proposed a scale-invariant Mask R-CNN model for robust and efficient pedestrian detection. Again, the proposed framework is also useful in suspicious activity recognition on recent benchmark databases. This well-organized comparison helps to identify problems and challenges in this domain. In the future, more experimentation is required for pose estimation and pedestrian trajectory identification and detection.

References

[1] Ahmed M., Jahangir M., Afzal H. (2015) “Using Crowd-source based features from social media and Conventional features to predict the movies popularity”, *IEEE International Conference on Smart Cities, Social Community and Sustained Community*, China, pp. 273–278.

[2] Bergmann P., Meinhardt T., Taixe L. (2019) “Tracking without bells and whistles”, *IEEE International Conference ICCV*, Seoul, Korea, pp. 1-16.

[3] Dollar P., Wojek C., Schiele B., and Perona P. (2012) “Pedestrian Detection: An Evaluation of the State of the Art”, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **34 (4)**: 743-761.

[4] Samsi S., Weiss M., Bestor D., Li D., Jones M., Reuther A., Edelman D., Arcand W. and Byun C. (2021) “The MIT Supercloud Dataset”, *International Cornell Journal of Distributed, Parallel, and Cluster Computing*, **2108 (02037)**: 1-10.

[5] Silberstein S., Levi D., Kogan V. and Gazit R. (2014) “Vision-based pedestrian detection for rear-view cameras”, *IEEE Intelligent Vehicles Symposium*, pp. 853-860, Dearborn, MI, USA.

[6] Alom M. and Taha T. (2017) “Robust multi-view pedestrian tracking using neural networks”, *IEEE National Conference on Aerospace and Electronics*, pp. 17-22, Dayton, OH, USA.

[7] Zhang X., Park S., Beeler T., Bradley D., Tang S., Hilliges O. (2020) “ETH Gaze: A Large-Scale Dataset for Gaze Estimation Under Extreme Head Pose and Gaze Variation” *European Conference on Computer Vision (ECCV),* Springer. Lecture Notes in Computer Science, pp. 1-10.

[8] Wojek C., Walk S., Schiele B., (2009) “Multiresolution model for Object Detection”, *IEEE Computer Vision and Pattern Recognition (CVPR)*, June 20-25, 2009, Miami, Florida, USA.

[9] Nguyen T., Soo K., (2013). “Fast Pedestrian Detection Using Histogram of Oriented Gradients and Principal Components Analysis”. *International Journal of Contents*, **2013 (1)**: 1-20.

[10] Everingham, M., Van L., Williams, C. (2010) “The Pascal Visual Object Classes (VOC) Challenge”, *International Journal of Computer Vision*, Springer, **88 (1)**: 303–338.

[11] Lin T. (2014), “Microsoft COCO: Common Objects in Context. ECCV 2014”, *Lecture Notes in Computer Science, Springe*r, **8693 (1)**: 740-755.

[12] Nicolai W., Bewley A., Dietrich P. (2017) “Simple online and real-time tracking with a deep association metric”, *IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649.

[13] Jifeng D., Yi L., Kaiming H., Sun J., (2016) “R-FCN: Object detection via region-based fully convolutional networks”, *IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 1-11.

[14] Everingham, M., Eslami, S., V. G., Williams C., Winn J. (2015) “The PASCAL VOC Challenge: A Retrospective”*, International Journal of Computer Vision (IJCV), Springer*, **11 (1)**: 98-136.

[15] Kaiming H., Georgia G., Dollar P. and Girshick R. (2020) “Mask R-CNN, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **42 (2)**: pp. 386-397.

[16] Viola P. and Jones M. (2001) “Rapid object detection using a boosted cascade of simple features,” *IEEE Computer Vision and Pattern Recognition (CVPR)*, USA, pp. I-I.

[17] Felzenszwalb P., Girshick R., McAllester D. and Ramanan D. (2010) “Object detection with discriminatively trained part- based models”, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **32 (9)**: 1627–1645.

[18] Dalal N. and Triggs B. (2005) “Histograms of oriented gradients for human detection”, *IEEE Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, pp. 886–893.

[19] Muhammad N., Hussain M., Muhammad G., Bebis G., (2011), “Copy-move forgery detection using dyadic wavelet transform,” *International Conference on Computer Graphics*, Singapore, pp. 103–108.

[20] Muhammad G., Hossain M., Kumar N., (2021) “EEG-based pathology detection for home health monitoring”, *IEEE Journal on Selected Areas in Community*, **39 (2)**: 603–610.

[21] Muhammad G., Alhamid M., and Long X., (2019) “Computing and processing on the edge: Smart pathology detection for connected healthcare”*, IEEE Network*, **33 (1)**: 44–49.

[22] Girshick R., Donahue J., Darrell T., and Malik J., (2014), “Rich feature hierarchies for accurate object detection and semantic segmentation” *IEEE Computer Vision and Pattern Recognition (CVPR),* Columbus, OH, USA, pp. 580–587.

[23] He K., Zhang X., Ren S., and Sun J., (2015) “Spatial pyramid pooling in deep convolutional networks for visual recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),* **37 (9)**: 1904–1916.

[24] Girshick R., (2015) “Fast R-CNN”, *IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1440–1448.

[25] Ren S., He K., Girshick R. and Sun J., (2016) “Faster R-CNN: Towards real-time object detection with region proposal networks”, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),* **39 (6)**: 1137–1149.

[26] He K., Gkioxari G., Dollar P., and Girshick R., (2017) “Mask R-CNN” *IEEE Computer Vision and Pattern Recognition (CVPR)*, Venice, Italy, pp. 2980–2988.

[27] Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., (2016) “SSD: Single shot multi-box detector” *European Conference on Computer Vision (ECCV)*, Cham, Springer, pp. 21–37.

[28] Redmon J., Girshick R. and Farhadi A., (2016) “You only look once: unified, real-time object detection”, *IEEE Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 779–788.

[29] Senst T. and Sikora T, (2018) “Extending IOU based multi-object tracking by visual information”, *IEEE International Conference on Advanced Video and Signal Based Surveillance*, Auckland, New Zealand, pp. 1-7.

[30] Barnardin K. and Stiefelhagen R. (2008) “Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics”, EURASIP *Journal on Image and Video Processing*, Springer, **1 (1)**: 246-309.

[31] Kratz L. and Nishino K. (2012) “Tracking Pedestrians Using Local Spatio-Temporal Motion Patterns in Extremely Crowded Scenes” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **34 (5)**: 987-1002.

[32] Wang S. and Miao Z. (2010) “Anomaly Detection in Crowd Scene”, *IEEE International Conference on Signal Processing*, pp. 1220-1223, Oct. 24-28, Beijing, China.

[33] Wang S., Miao Z. (2010) “Anomaly Detection in Crowd Scene Using Historical Information”, *IEEE Intelligence Signal Processing and Communication System*, pp. 1-4.